Chapter 4

Parameter Estimation

Thus far we have concerned ourselves primarily with *probability theory*: what events may occur with what probabilities, given a model family and choices for the parameters. This is useful only in the case where we know the precise model family and parameter values for the situation of interest. But this is the exception, not the rule, for both scientific inquiry and human learning & inference. Most of the time, we are in the situation of processing data whose generative source we are uncertain about. In Chapter 2 we briefly covered elementary density estimation, using relative-frequency estimation, histograms and kernel density estimation. In this chapter we delve more deeply into the theory of probability density estimation, focusing on inference within parametric families of probability distributions (see discussion in Section 2.11.2). We start with some important properties of estimators, then turn to basic frequentist parameter estimation (maximum-likelihood estimation and corrections for bias), and finally basic Bayesian parameter estimation.

4.1 Introduction

Consider the situation of the first exposure of a native speaker of American English to an English variety with which she has no experience (e.g., Singaporean English), and the problem of inferring the probability of use of active versus passive voice in this variety with a simple transitive verb such as *hit*:

(1) The ball hit the window. (Active)

(2) The window was hit by the ball. (Passive)

There is ample evidence that this probability is contingent on a number of features of the utterance and discourse context (e.g., Weiner and Labov, 1983), and in Chapter 6 we cover how to construct such richer models, but for the moment we simplify the problem by assuming that active/passive variation can be modeled with a binomial distribution (Section 3.4) with parameter π characterizing the probability that a given potentially transitive clause eligible for passivization will in fact be realized as a passive.[1] The question faced by the native American English speaker is thus, what inferences should we make about π on the basis of limited exposure to the new variety? This is the problem of PARAMETER ESTIMATION, and it is a central part of statistical inference. There are many different techniques for parameter estimation; any given technique is called an ESTIMATOR, which is applied to a set of data to construct an estimate. Let us briefly consider two simple estimators for our example.

Estimator 1. Suppose that our American English speaker has been exposed to n transitive sentences of the variety, and m of them have been realized in the passive voice in eligible clauses. A natural estimate of the binomial parameter π would be m/n. Because m/n is the relative frequency of the passive voice, this is known as the RELATIVE FREQUENCY ESTIMATE (RFE; see Section 2.11.1). In addition to being intuitive, we will see in Section 4.3.1 that the RFE can be derived from deep and general principles of optimality in estimation procedures. However, RFE also has weaknesses. For instance, it makes no use of the speaker's knowledge of her native English variety. In addition, when n is small, the RFE is unreliable: imagine, for example, trying to estimate π from only two or three sentences from the new variety.

Estimator 2. Our speaker presumably knows the probability of a passive in American English; call this probability q. An extremely simple estimate of π would be to ignore all new evidence and set $\pi = q$, regardless of how much data she has on the new variety. Although this option may not be as intuitive as Estimator 1, it has certain advantages: it is extremely reliable and, if the new variety is not too different from American English, reasonably accurate as well. On the other hand, once the speaker has had considerable exposure to the new variety, this approach will almost certainly be inferior to relative frequency estimation. (See Exercise to be included with this chapter.)

In light of this example, Section 4.2 describes how to assess the quality of an estimator in conceptually intuitive yet mathematically precise terms. In Section 4.3, we cover FREQUENTIST approaches to parameter estimation, which involve procedures for constructing point estimates of parameters. In particular we focus on maximum-likelihood estimation and close variants, which for multinomial data turns out to be equivalent to Estimator 1 above.In Section 4.4, we cover BAYESIAN approaches to parameter estimation, which involve placing probability distributions over the range of possible parameter values. The Bayesian estimation technique we will cover can be thought of as intermediate between Estimators 1 and 2.

4.2 Desirable properties for estimators

In this section we briefly cover three key properties of any estimator, and discuss the desirability of these properties.

[1] By this probability we implicitly conditionalize on the use of a transitive verb that is eligible for passivization, excluding intransitives and also unpassivizable verbs such as *weigh*.

4.2.1 Consistency

An estimator is CONSISTENT if the estimate $\hat{\theta}$ it constructs is guaranteed to converge to the true parameter value θ as the quantity of data to which it is applied increases. Figure 4.1 demonstrates that Estimator 1 in our example is consistent: as the sample size increases, the probability that the relative-frequency estimate $\hat{\pi}$ falls into a narrow band around the true parameter π grows asymptotically toward 1 (this behavior can also be proved rigorously; see Section 4.3.1). Estimator 2, on the other hand, is not consistent (so long as the American English parameter q differs from π), because it ignores the data completely. Consistency is nearly always a desirable property for a statistical estimator.

4.2.2 Bias

If we view the collection (or *sampling*) of data from which to estimate a population parameter as a stochastic process, then the parameter estimate $\hat{\theta}_\eta$ resulting from applying a pre-determined estimator η to the resulting data can be viewed as a continuous random variable (Section 3.1). As with any random variable, we can take its expectation. In general, it is intuitively desirable that the expected value of the estimate be equal (or at least close) to the true parameter value θ, but this will not always be the case. The BIAS of an estimator η is defined as the deviation of the expectation from the true value: $E[\hat{\theta}_\eta] - \theta$. All else being equal, the smaller the bias in an estimator the more preferable. An estimator for which the bias is zero—that is, $E[\theta_\eta] = \theta$—is called UNBIASED.

Is Estimator 1 in our passive-voice example biased? The relative-frequency estimate $\hat{\pi}$ is $\frac{m}{n}$, so $E[\hat{\pi} = E[\frac{m}{n}]$. Since n is fixed, we can move it outside of the expectation (see linearity of the expectation in Section 3.3.1) to get

$$E[\hat{\pi}] = \frac{1}{n} E[m]$$

But m is just the number of passive-voice utterances heard, and since m is binomially distributed, $E[m] = \pi n$. This means that

$$\begin{aligned} E[\hat{\pi}] &= \frac{1}{n} \pi n \\ &= \pi \end{aligned}$$

So Estimator 1 is unbiased. Estimator 2, on the other hand, has bias $q - \pi$.

4.2.3 Variance (and efficiency)

Suppose that our speaker has decided to use Estimator 1 to estimate the probability π of a passive, and has been exposed to n utterances. The intuition is extremely strong that she should use *all* n utterances to form her relative-frequency estimate $\hat{\pi}$, rather than, say, using only the first $n/2$. But why is this the case? Regardless of how many utterances she uses with Estimator 1, her estimate will be unbiased (think about this carefully if you are not immediately convinced). But our intuitions suggest that an estimate using less data is less reliable: it is likely to vary more dramatically due to pure freaks of chance.

It is useful to quantify this notion of reliability using a natural statistical metric: the VARIANCE of the estimator, $\mathrm{Var}(\hat{\theta})$ (Section 4.2.3). All else being equal, an estimator with smaller variance is preferable to one with greater variance. This idea, combined with a bit more simple algebra, quantitatively explains the intuition that more data is better for Estimator 1:

$$\begin{aligned}
\mathrm{Var}(\hat{\pi}) &= \mathrm{Var}\left(\frac{m}{n}\right) \\
&= \frac{1}{n^2}\mathrm{Var}(m) \qquad \text{(From scaling a random variable, Section 3.3.3)}
\end{aligned}$$

Since m is binomially distributed, and the variance of the binomial distribution is $n\pi(1-\pi)$ (Section 3.4), so we have

$$\mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n} \tag{4.1}$$

So variance is inversely proportional to the sample size n, which means that relative frequency estimation is more reliable when used with larger samples, consistent with intuition.

It is almost always the case that each of bias and variance comes at the cost of the other. This leads to what is sometimes called BIAS-VARIANCE TRADEOFF: one's choice of estimator may depend on the relative importance of expected accuracy versus reliability in the task at hand. The bias-variance tradeoff is very clear in our example. Estimator 1 is unbiased, but has variance that can be quite high when samples size n is small. Estimator 2 is biased, but it has zero variance. Which of the two estimators is preferable is likely to depend on the sample size. If our speaker anticipates that she will have very few examples of transitive sentences in the new English variety to go on, and also anticipates that the new variety will not be hugely different from American English, she may well prefer (and with good reason) the small bias of Estimator 2 to the large variance of Estimator 1. The lower-variance of two estimators is called the more EFFICIENT estimator, and the EFFICIENCY of one estimator η_1 relative to another estimator η_2 is the ratio of their variances, $\mathrm{Var}(\hat{\theta}_{\eta_1})/\mathrm{Var}(\hat{\theta}_{\eta_2})$.

4.3 Frequentist parameter estimation and prediction

We have just covered a simple example of parameter estimation and discussed key properties of estimators, but the estimators we covered were (while intuitive) given no theoretical underpinning. In the remainder of this chapter, we will cover a few major mathematically motivated estimation techniques of general utility. This section covers FREQUENTIST estimation techniques. In frequentist statistics, an estimator gives a point estimate for the parameter(s) of interest, and estimators are preferred or dispreferred on the basis of their general behavior, notably with respect to the properties of consistency, bias, and variance discussed in Section 4.2. We start with the most widely-used estimation technique, MAXIMUM-LIKELIHOOD ESTIMATION.

4.3.1 Maximum Likelihood Estimation

We encountered the notion of the LIKELIHOOD in Chapter 2, a basic measure of the quality of a set of predictions with respect to observed data. In the context of parameter estimation, the likelihood is naturally viewed as a function of the parameters $\boldsymbol{\theta}$ to be estimated, and is defined as in Equation (2.29)—the joint probability of a set of observations, conditioned on a choice for $\boldsymbol{\theta}$—repeated here:

$$\text{Lik}(\boldsymbol{\theta};\boldsymbol{y}) \equiv P(\boldsymbol{y}|\boldsymbol{\theta}) \tag{4.2}$$

Since good predictions are better, a natural approach to parameter estimation is to choose the set of parameter values that yields the best predictions—that is, the parameter that *maximizes the likelihood* of the observed data. This value is called the MAXIMUM LIKELIHOOD ESTIMATE (MLE), defined formally as:[2]

$$\hat{\boldsymbol{\theta}}_{MLE} \overset{\text{def}}{=} \arg\max_{\boldsymbol{\theta}} \text{Lik}(\boldsymbol{\theta};\boldsymbol{y}) \tag{4.3}$$

In nearly all cases, the MLE is consistent (Cramer, 1946), and gives intuitive results. In many common cases, it is also unbiased. For estimation of multinomial probabilities, the MLE also turns out to be the relative-frequency estimate. Figure 4.2 visualizes an example of this. The MLE is also an intuitive and unbiased estimator for the means of normal and Poisson distributions.

Likelihood as function of data or model parameters?

In Equation (4.2) I defined the likelihood as a function first and foremost of the parameters of one's model. I did so as

4.3.2 Limitations of the MLE: variance

As intuitive and general-purpose as it may be, the MLE has several important limitations, hence there is more to statistics than maximum-likelihood. Although the MLE for multinomial distributions is unbiased, its variance is problematic for estimating parameters that determine probabilities of events with low expected counts. This can be a major problem

[2] The expression $\arg\max_x f(x)$ is defined as "the value of x that yields the maximum value for the expression $f(x)$." It can be read as "arg-max over x of $f(x)$."

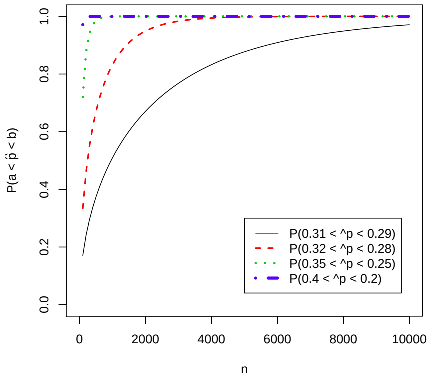


Figure 4.1: Consistency of relative frequency estimation. Plot indicates the probability with which the relative-frequency estimate $\hat{\pi}$ for binomial distribution with parameter $\pi = 0.3$ lies in narrrow ranges around the true parameter value as a function of sample size n.

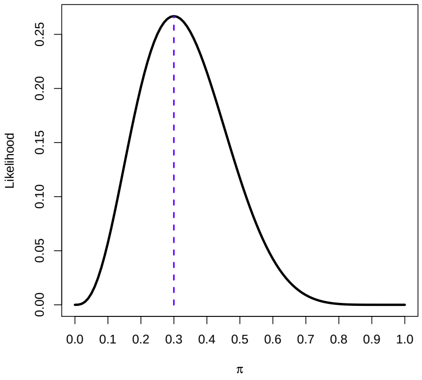


Figure 4.2: The likelihood function for the binomial parameter π for observed data where $n = 10$ and $m = 3$. The MLE is the RFE for the binomial distribution. Note that this graph is *not* a probability density and the area under the curve is much less than 1.

even when the sample size is very large. For example, WORD N-GRAM PROBABILITIES—the probability distribution over the next word in a text given the previous $n-1$ words of context—are of major interest today not only in applied settings such as speech recognition but also in the context of theoretical questions regarding language production, comprehension, and acquisition (e.g., Gahl, 2008; Saffran et al., 1996b; 2-gram probabilities are sometimes called TRANSITIONAL PROBABILITIES). N-gram probability models are simply collections of large multinomial distributions (one distribution per context). Yet even for extremely high-frequency preceding contexts, such as the word sequence *near the*, there will be many possible next words that are improbable yet not impossible (for example, *reportedly*). Any word that does not appear in the observed data in that context will be assigned a conditional probability of zero by the MLE. In a typical n-gram model there will be many, many such words—the problem of DATA SPARSITY. This means that the MLE is a terrible means of prediction for n-gram word models, because if *any* unseen word continuation appears in a new dataset, the MLE will assign zero likelihood to the *entire dataset*. For this reason, there is a substantial literature on learning high-quality n-gram models, all of which can in a sense be viewed as managing the variance of estimators for these models while keeping the bias reasonably low (see Chen and Goodman, 1998 for a classic survey).

4.3.3 Limitations of the MLE: bias

In addition to these problems with variance, the MLE is biased for some types of model parameters. Imagine a linguist interested in inferring the original time of introduction of a novel linguistic expression currently in use today, such as the increasingly familiar phrase *the boss of me*, as in:[3]

(3) "You're too cheeky," said Astor, sticking out his tongue. "You're not the boss of me." (Tool, 1949, cited in *Language Log* by Benjamin Zimmer, 18 October 2007)

The only direct evidence for such expressions is, of course, attestations in written or recorded spoken language. Suppose that the linguist had collected 60 attestations of the expression, the oldest of which was recored 120 years ago.

From a probabilistic point of view, this problem involves choosing a probabilistic model whose generated observations are n attestation dates $\boldsymbol{y}$ of the linguistic expression, and one of whose parameters is the earliest time at which the expression is coined, or t_0. When the problem is framed this way, the linguist's problem is to devise a procedure for constructing a parameter estimate $\hat{t}_0$ from observations. For expository purposes, let us oversimplify and use the uniform distribution as a model of how attestation dates are generated.[4] Since the innovation is still in use today (time t_{now}), the parameters of the uniform distribution are $[t_0, t_{now}]$ and the only parameter that needs to be estimated is t_{now}. Let us arrange our attestation dates in chronological order so that the earliest date is y_1.

What is the maximum-likelihood estimate $\hat{t}_0$? For a given choice of t_0, a given date y_i either falls in the interval $[t_0, t_{now}]$ or it does not. From the definition of the uniform distribution (Section 2.7.1) we have:

$$P(y_i|t_0, t_{now}) = \begin{cases} \frac{1}{t_{now}-t_0} & t_0 \le y_i \le t_{now} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Due to independence, the likelihood for the interval boundaries is $\text{Lik}(t_0) = \prod_i P(y_i|t_0, t_{now})$. This means that for any choice of interval boundaries, if at least one date lies before t_0, the entire likelihood is zero! Hence the likelihood is non-zero only for interval boundaries containing all dates. For such boundaries, the likelihood is

$$\begin{aligned} \text{Lik}(t_0) &= \prod_{i=1}^{n} \frac{1}{t_{now} - t_0} && (4.5) \\ &= \frac{1}{(t_{now} - t_0)^n} && (4.6) \end{aligned}$$

This likelihood grows larger as $t_{now} - t_0$ grows smaller, so it will be maximized when the interval length $t_{now} - t_0$ is as short as possible—namely, when t_0 is set to the earliest attested

[3] This phrase has been the topic of intermittent discussion on the *Language Log* blog since 2007.

[4] This is a dramatic oversimplification, as it is well known that linguistic innovations prominent enough for us to notice today often followed an S-shaped trajectory of usage frequency (Bailey, 1973; Cavall-Sforza and Feldman, 1981; Kroch, 1989; Wang and Minett, 2005). However, the general issue of bias in maximum-likelihood estimation present in the oversimplified uniform-distribution model here also carries over to more complex models of the diffusion of linguistic innovations.

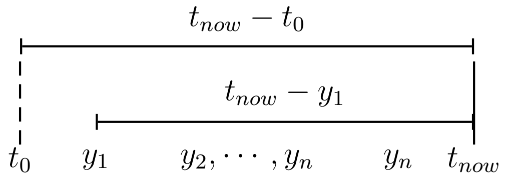


Figure 4.3: The bias of the MLE for uniform distributions

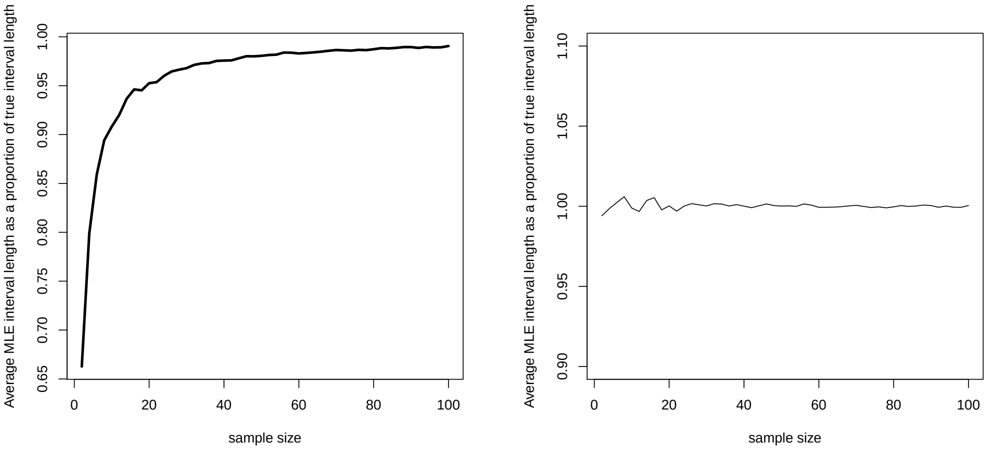


Figure 4.4: Bias of the MLE (left) and the bias-corrected estimator (right), shown numerically using 500 simulations for each value of n.

date y_1. This fact is illustrated in Figure 4.3: the tighter the posited interval between t_0 and t_{now}, the greater the resulting likelihood.

You probably have the intuition that this estimate of the contact interval duration is conservative: certainly the novel form appeared in English no later than t_0, but it seems rather unlikely that the first use in the language was also the first attested use![5] This intuition is correct, and its mathematical realization is that the MLE for interval boundaries of a uniform distribution is biased. Figure 4.4 visualizes this bias in terms of average interval length (over a number of samples) as a function of sample size.

For any finite sample size, the MLE is biased to underestimate true interval length, although this bias decreases as sample size increases (as well it should, because the MLE is a consistent estimator). Fortunately, the size of the MLE's bias can be quantified analytically: the expected ML-estimated interval size is $\frac{n}{n+1}$ times the true interval size.Therefore, if we adjust the MLE by multiplying it by $\frac{n+1}{n}$, we arrive at an unbiased estimator for interval length. The correctness of this adjustment is confirmed by the right-hand plot in Figure 4.4. In the case of our historical linguist with three recovered documents, we achieve the estimate

[5] The intuition may be different if the first attested use was by an author who is known to have introduced a large number of novel expressions into the language which subsequently gained in popularity. This type of situation would point to a need for a more sophisticated probabilistic model of innovation, diffusion, and attestation.

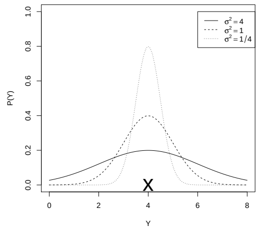


Figure 4.5: Bias in the MLE for σ of a normal distribution.

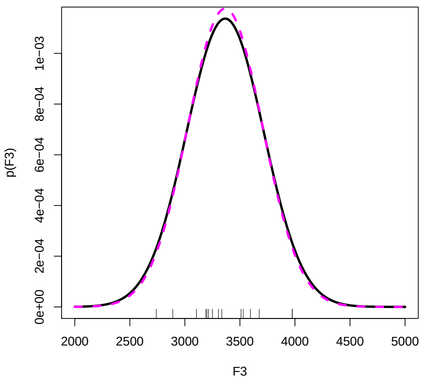


Figure 4.6: Point estimation of a normal distribution. The maximum-likelihood estimate is the dotted magenta line; the bias-adjusted estimate is solid black.

$$\hat{t}_0 = 120 \times \frac{61}{60} = 122 \text{ years ago}$$

Furthermore, there is a degree of intuitiveness about the behavior of the adjustment in extreme cases: if $N = 1$, the adjustment would be infinite, which makes sense: one cannot estimate the size of an unconstrained interval from a single observation.

Another famous example of bias in the MLE is in estimating the variance of a normal distribution. The MLEs for mean and variance of a normal distribution as estimated from a set of N observations $\boldsymbol{y}$ are as follows:

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_i y_i \qquad \text{(i.e. the sample mean)} \qquad (4.7)$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_i (y_i - \hat{\mu})^2 \qquad \text{(i.e. the sample variance divided by } N\text{)} \qquad (4.8)$$

While it turns out that $\hat{\mu}_{MLE}$ is unbiased, $\hat{\sigma}^2_{MLE}$ is biased for reasons similar to those given for interval size in the uniform distribution. You can see this graphically by imagining the MLE for a single observation, as in Figure 4.5. As $\hat{\sigma}^2$ shrinks, the likelihood of the observation will continue to rise, so that the MLE will push the estimated variance to be arbitrarily small. This is a type of OVERFITTING (see Section 2.11.5).

It turns out that the this bias can be eliminated by adjusting the MLE by the factor $\frac{N}{N-1}$. This adjusted estimate of σ^2 is called S^2:

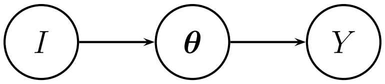


Figure 4.7: The structure of a simple Bayesian model. Observable data Y and prior beliefs I are conditionally independent given the model parameters.

$$S^2 = \frac{N}{N-1}\hat{\sigma}^2_{MLE} \tag{4.9}$$

$$= \frac{1}{N-1}\sum_i (y_i - \hat{\mu})^2 \tag{4.10}$$

This is the most frequently used estimate of the underlying variance of a normal distribution from a sample. In `R`, for example, the function `var()`, which is used to obtain sample variance, computes S^2 rather than $\hat{\sigma}_{MLE}$. An example of estimating normal densities is shown in Figure 4.6, using F3 formants from 15 native English-speaking children on the vowel [æ]. The MLE density estimate is a slightly narrower curve than the bias-adjusted estimate.

4.4 Bayesian parameter estimation and density estimation

In frequentist statistics as we have discussed thus far, one uses observed data to construct a point estimate for each model parameter. The MLE and bias-adjusted version of the MLE are examples of this. In Bayesian statistics, on the other hand, parameter estimation involves placing a *probability distribution* over model parameters. In fact, there is no conceptual difference between parameter estimation (inferences about $\boldsymbol{\theta}$) and prediction or density estimation (inferences about future $\boldsymbol{y}$) in Bayesian statistics.

4.4.1 Anatomy of inference in a simple Bayesian model

A simple Bayesian model has three components. Observable data are generated as random variables $\boldsymbol{y}$ in some model from a model family with parameters $\boldsymbol{\theta}$. Prior to observing a particular set of data, however, we already have beliefs/expectations about the possible model parameters $\boldsymbol{\theta}$; we call these beliefs I. These beliefs affect $\boldsymbol{y}$ only through the mediation of the model parameters—that is, $\boldsymbol{y}$ and I are *conditionally independent* given $\boldsymbol{\theta}$ (see Section 2.4.2). This situation is illustrated in Figure 6.1, which has a formal interpretation as a graphical model (Appendix C).

In the Bayesian framework, both parameter estimation and density estimation simply involve the application of Bayes' rule (Equation (2.5)). For example, parameter estimation means calculating the probability distribution over $\boldsymbol{\theta}$ given observed data $\boldsymbol{y}$ and our prior beliefs I. We can use Bayes rule to write this distribution as follows:

$$P(\boldsymbol{\theta}|\boldsymbol{y}, I) = \frac{P(\boldsymbol{y}|\theta, I)P(\theta|I)}{P(\boldsymbol{y}|I)} \tag{4.11}$$

$$= \frac{\overbrace{P(\boldsymbol{y}|\boldsymbol{\theta})}^{\text{Likelihood for } \boldsymbol{\theta}} \qquad \overbrace{P(\boldsymbol{\theta}|I)}^{\text{Prior over } \boldsymbol{\theta}}}{\underbrace{P(\boldsymbol{y}|I)}_{\text{Likelihood marginalized over } \boldsymbol{\theta}}} \qquad (\text{because } \boldsymbol{y} \perp I \mid \boldsymbol{\theta}) \tag{4.12}$$

The numerator in Equation (4.12) is composed of two quantities. The first term, $P(\boldsymbol{y}|\boldsymbol{\theta})$, should be familiar from Section 2.11.5: it is the likelihood of the parameters $\boldsymbol{\theta}$ for the data $\boldsymbol{y}$. As in much of frequentist statistics, the likelihood plays a central role in parameter estimation in Bayesian statistics. However, there is also a second term, $P(\boldsymbol{\theta}|I)$, the PRIOR DISTRIBUTION over $\boldsymbol{\theta}$ given only I. The complete quantity (4.12) is the POSTERIOR DISTRIBUTION over $\boldsymbol{\theta}$. It is important to realize that the terms "prior" and "posterior" in no way imply any temporal ordering on the realization of different events. The only thing that $P(\boldsymbol{\theta}|I)$ is "prior" to is the incorporation of the particular dataset $\boldsymbol{y}$ into inferences about $\boldsymbol{\theta}$. I can in principle incorporate all sorts of knowledge, including other data sources, scientific intuitions, or—in the context of language acquisition—innate biases. Finally, the denominator is simply the MARGINAL LIKELIHOOD $P(\boldsymbol{y}|I) = \int_{\boldsymbol{\theta}} P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}|I)\, d\boldsymbol{\theta}$ (it is the model parameters $\boldsymbol{\theta}$ that are being marginalized over; see Section 3.2). The data likelihood is often the most difficult term to calculate, but in many cases its calculation can be ignored or circumvented because we can accomplish everything we need by computing posterior distributions up to a normalizing constant (Section 2.8; we will see an new example of this in the next section).

Since Bayesian inference involves placing probability distributions on model parameters, it becomes useful to work with probability distributions that are specialized for this purpose. Before we move on to our first simple example of Bayesian parameter and density estimation, we'll now introduce one of the simplest (and most easily interpretable) such probability distributions: the beta distribution.

4.4.2 The beta distribution

The BETA DISTRIBUTION is important in Bayesian statistics involving binomial distributions. It has two parameters α_1, α_2 and is defined as follows:

$$P(\pi|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)}\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1} \qquad (0 \leq \pi \leq 1, \alpha_1 > 0, \alpha_2 > 0) \tag{4.13}$$

where the BETA FUNCTION $B(\alpha_1, \alpha_2)$ (Section B.1) serves as a normalizing constant:

$$B(\alpha_1, \alpha_2) = \int_0^1 \pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}\, d\pi \tag{4.14}$$

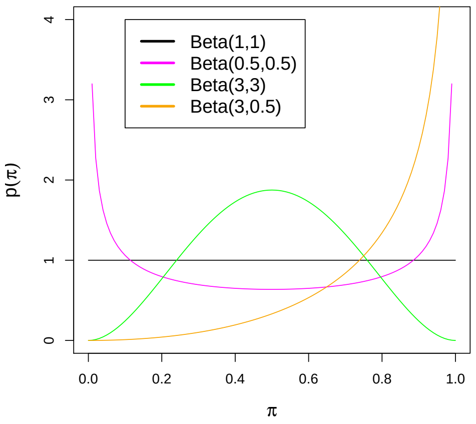


Figure 4.8: Beta distributions

Figure 4.8 gives a few examples of beta densities for different parameter choices. The beta distribution has a mean of $\frac{\alpha_1}{\alpha_1+\alpha_2}$ and mode (when both $\alpha_1, \alpha_2 > 1$) of $\frac{\alpha_1-1}{\alpha_1+\alpha_2-2}$. Note that a uniform distribution on $[0,1]$ results when $\alpha_1 = \alpha_2 = 1$.

Beta distributions and beta functions are very often useful when dealing with Bayesian inference on binomially-distributed data. One often finds oneself in the situation of knowing that some random variable X is distributed such that $P(X) \propto \pi^a(1-\pi)^b$, but not knowing the normalization constant. If and when you find yourself in this situation, recognize that X must be beta-distributed, which allows you to determine the normalization constant immediately. Additionally, whenever one is confronted with an integral of the form $\int_0^1 \pi^a(1-\pi)^b \, d\pi$ (as in Section 5.2.1), recognize that it is a beta function, which will allow you to compute the integral very easily.

4.4.3 Simple example of Bayesian estimation with the binomial distribution

Historically, one of the major reasons that Bayesian inference has been avoided is that it can be computationally intensive under many circumstances. The rapid improvements in available computing power over the past few decades are, however, helping overcome this obstacle, and Bayesian techniques are becoming more widespread both in practical statistical applications and in theoretical approaches to modeling human cognition. We will see examples of more computationally intensive techniques later in the book, but to give the flavor of the Bayesian approach, let us revisit the example of our native American English speaker and her quest for an estimator for π, the probability of the passive voice, which turns out to be analyzable without much computation at all.

We have already established that transitive sentences in the new variety can be modeled using a binomial distribution where the parameter π characterizes the probability that a given transitive sentence will be in the passive voice. For Bayesian statistics, we must first specify the beliefs I that characterize the prior distribution $P(\boldsymbol{\theta}|I)$ to be held before any data from the new English variety is incorporated. In principle, we could use any proper probability distribution on the interval $[0,1]$ for this purpose, but here we will use the beta distribution (Section 4.4.2). In our case, specifying prior knowledge I amounts to choosing beta distribution parameters α_1 and α_2.

Once we have determined the prior distribution, we are in a position to use a set of observations $\boldsymbol{y}$ to do parameter estimation. Suppose that the observations $\boldsymbol{y}$ that our speaker has observed are comprised of n total transitive sentences, m of which are passivized. Let us simply instantiate Equation (4.12) for our particular problem:

$$P(\pi|\boldsymbol{y},\alpha_1,\alpha_2) = \frac{P(\boldsymbol{y}|\pi)P(\pi|\alpha_1,\alpha_2)}{P(\boldsymbol{y}|\alpha_1,\alpha_2)} \tag{4.15}$$

The first thing to notice here is that the denominator, $P(\boldsymbol{y}|\alpha_1,\alpha_2)$, is not a function of π. That means that it is a normalizing constant (Section 2.8). As noted in Section 4.4, we can often do everything we need without computing the normalizing constant, here we ignore the denominator by re-expressing Equation (4.15) in terms of proportionality:

$$P(\pi|\boldsymbol{y},\alpha_1,\alpha_2) \propto P(\boldsymbol{y}|\pi)P(\pi|\alpha_1,\alpha_2)$$

From what we know about the binomial distribution, the likelihood is $P(\boldsymbol{y}|\pi) = \binom{n}{m}\pi^m(1-\pi)^{n-m}$, and from what we know about the beta distribution, the prior is $P(\pi|\alpha_1,\alpha_2) = \frac{1}{B(\alpha_1,\alpha_2)}\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}$. Neither $\binom{n}{m}$ nor $B(\alpha_1,\alpha_2)$ is a function of π, so we can also ignore them, giving us

$$\begin{aligned} P(\pi|\boldsymbol{y},\alpha_1,\alpha_2) &\propto \overbrace{\pi^m(1-\pi)^{n-m}}^{\text{Likelihood}}\overbrace{\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}}^{\text{Prior}} \\ &\propto \pi^{m+\alpha_1-1}(1-\pi)^{n-m+\alpha_2-1} \end{aligned} \tag{4.16}$$

Now we can crucially notice that the posterior distribution on π itself has the form of a beta distribution (Equation (4.13)), with parameters α_1+m and α_2+n-m. This fact that the posterior has the same functional form as the prior is called CONJUGACY; the beta distribution is said to be CONJUGATE TO the binomial distribution. Due to conjugacy, we can circumvent the work of directly calculating the normalizing constant for Equation (4.16), and recover it from what we know about beta distributions. This gives us a normalizing constant of $B(\alpha_1+m,\alpha_2+n-m)$.

Now let us see how our American English speaker might apply Bayesian inference to estimating the probability of passivization in the new English variety. A reasonable prior distribution might involve assuming that the new variety could be somewhat like American English. Approximately 8% of spoken American English sentences with simple transitive

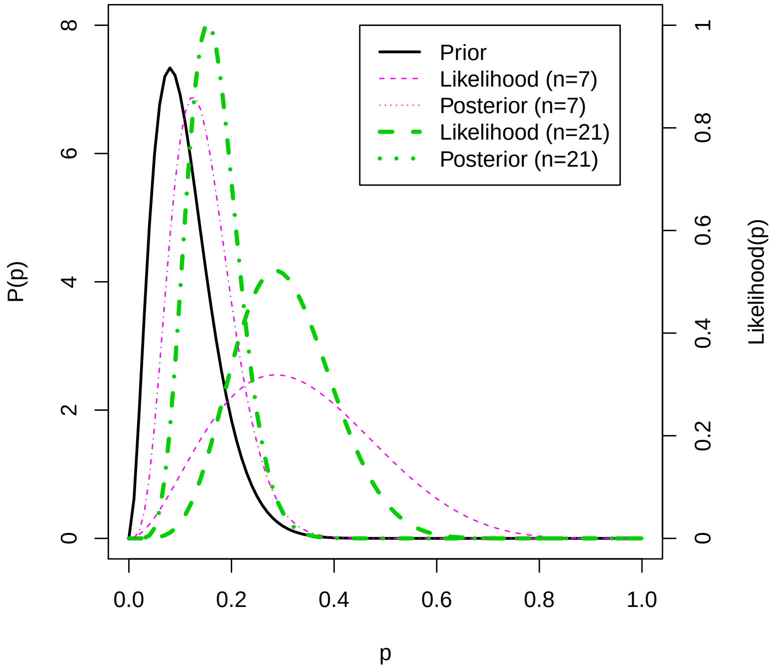


Figure 4.9: Prior, likelihood, and posterior distributions over π. Note that the likelihood has been rescaled to the scale of the prior and posterior; the original scale of the likelihood is shown on the axis on the right.

verbs are passives (Roland et al., 2007), hence our speaker might choose α_1 and α_2 such that the mode of $P(\pi|\alpha_1, \alpha_2)$ is near 0.08. A beta distribution has a mode if $\alpha_1, \alpha_2 > 1$, in which case the mode is $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$, so a reasonable choice might be $\alpha_1 = 3, \alpha_2 = 24$, which puts the mode of the prior distribution at $\frac{2}{25} = 0.08$.[6] Now suppose that our speaker is exposed to $n = 7$ transitive verbs in the new variety, and two are passivized $(m = 2)$. The posterior distribution will then be beta-distributed with $\alpha_1 = 3 + 2 = 5, \alpha_2 = 24 + 5 = 29$. Figure 4.9 shows the prior distribution, likelihood, and posterior distribution for this case, and also for the case where the speaker has been exposed to three times as much data in similar proportions $(n = 21, m = 6)$. In the $n = 7$, because the speaker has seen relatively little data, the prior distribution is considerably more peaked than the likelihood, and the posterior distribution is fairly close to the prior. However, as our speaker sees more and more data, the likelihood becomes increasingly peaked, and will eventually dominate in the behavior of the posterior (See Exercise to be included with this chapter).

In many cases it is useful to summarize the posterior distribution into a point estimate of the model parameters. Two commonly used such point estimates are the *mode* (which

[6] Compare with Section 4.3.1—the binomial likelihood function has the same shape as a beta distribution!

we covered a moment ago) and the *mean*. For our example, the posterior mode is $\frac{4}{32}$, or 0.125. Selecting the mode of the posterior distribution goes by the name of MAXIMUM A POSTERIORI (MAP) estimation. The mean of a beta distribution is $\frac{\alpha_1}{\alpha_1+\alpha_2}$, so our POSTERIOR MEAN is $\frac{5}{34}$, or about 0.15. There are no particular deep mathematical principles motivating the superiority of the mode over the mean or vice versa, although the mean should generally be avoided in cases where the posterior distribution is multimodal. The most "principled" approach to Bayesian parameter estimation is in fact *not* to choose a point estimate for model parameters after observing data, but rather to make use of the entire posterior distribution in further statistical inference.

Bayesian density estimation

The role played in density estimation by parameter estimation up to this point has been as follows: an estimator is applied to observed data to obtain an estimate for model parameters $\hat{\boldsymbol{\theta}}$, and the resulting probabilistic model determines a set of predictions for future data, namely the distribution $P(Y|\hat{\boldsymbol{\theta}})$. If we use Bayesian inference to form a posterior distribution on $\boldsymbol{\theta}$ and then summarize that distribution into a point estimate, we can use that point estimate in exactly the same way. In this sense, using a given prior distribution together with the MAP or posterior mean can be thought of as simply one more estimator. In fact, this view creates a deep connection between Bayesian inference and maximum-likelihood estimation: maximum-likelihood estimation (Equation (4.3)) is simply Bayesian MAP estimation when the prior distribution $P(\boldsymbol{\theta}|I)$ (Equation (4.11)) is taken to be uniform over all values of $\boldsymbol{\theta}$.

However, in the purest Bayesian view, it is undesirable to summarize our beliefs about model parameters into a point estimate, because this discards information. In Figure 4.9, for example, the two likelihoods are peaked at the same place, but the $n = 21$ likelihood is more peaked than the $n = 7$ likelihood. This translates into more peakedness and therefore more certainty in the posterior; this certainty is not reflected in the MLE or even in the MAP estimate. Pure Bayesian density estimation involves *marginalization* (Section 3.2) over the model parameters, a process which automatically incorporates this degree of certainty. That is, we estimate a density over new observations $\boldsymbol{y}_{new}$ as:

$$P(\boldsymbol{y}_{new}|\boldsymbol{y}, I) = \int_{\boldsymbol{\theta}} P(\boldsymbol{y}_{new}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{y}, I)\, d\boldsymbol{\theta} \tag{4.17}$$

where $P(\boldsymbol{\theta}|\boldsymbol{y}, I)$ is familiar from Equation (4.12).

Suppose, for example, that after hearing her n examples from the new English dialect, our speaker wanted to predict the number of passives r she would hear after the next k trials. We would have:

$$P(r|k, I, \boldsymbol{y}) = \int_0^1 P(r|k, \pi) P(\pi|\boldsymbol{y}, I)\, d\pi$$

This expression can be reduced to

$$P(r|k, I, \boldsymbol{y}) = \binom{k}{r} \frac{\prod_{i=0}^{r-1}(\alpha_1 + m + i) \prod_{i=0}^{k-r-1}(\alpha_2 + n - m + i)}{\prod_{i=0}^{k-1}(\alpha_1 + \alpha_2 + n + i)} \tag{4.18}$$

$$= \binom{k}{r} \frac{B(\alpha_1 + m + r, \alpha_2 + n - m + k - r)}{B(\alpha_1 + m, \alpha_2 + n - m)} \tag{4.19}$$

which is an instance of what is known as the BETA-BINOMIAL MODEL. The expression may seem formidable, but experimenting with specific values for k and r reveals that it is simpler than it may seem. For a single trial ($k = 1$), for example, this expression reduces to $P(r = 1|k, I, \boldsymbol{y}) = \frac{\alpha_1+m}{\alpha_1+\alpha_2+n}$, which is exactly what would be obtained by using the posterior mean. For two trials ($k = 2$), we would have $P(r = 1|k, I, \boldsymbol{y}) = 2\frac{(\alpha_1+m)(\alpha_2+n-m)}{(\alpha_1+\alpha_2+n)(\alpha_1+\alpha_2+n+1)}$, which is slightly less than what would be obtained by using the posterior mean.[7] This probability mass lost from the $r = 1$ outcome is redistributed into the more extreme $r = 0$ and $r = 2$ outcomes. For $k > 1$ trials in general, the beta-binomial model leads to density estimates of greater variance—also called DISPERSION in the modeling context—than for the binomial model using posterior mean. This is illustrated in Figure 4.10. The reason for this greater dispersion is that different future trials are only conditionally independent given a fixed choice of the binomial parameter π. Because there is residual uncertainty about this parameter, successes on different future trials are positively correlated in the Bayesian prediction despite the fact that they are conditionally independent given the underlying model parameter (see also Section 2.4.2 and Exercise 2.2). This is an important property of a wide variety of models which involve marginalization over intermediate variables (in this case the binomial parameter); we will return to this in Chapter 8 and later in the book.

4.5 Computing approximate Bayesian inferences with sampling techniques

In the example of Bayesian inference given in Section 4.4.3, we were able to express both (i) the posterior probability over the binomial parameter π, and (ii) the probability distribution over new observations as the CLOSED-FORM expressions[8] shown in Equations (4.16)

[7] With the posterior mean, the term $(\alpha_1 + \alpha_2 + n + 1)$ in the denominator would be replaced by another instance of $(\alpha_1 + \alpha_2 + n)$, giving us

$$P(r = 1|k, \hat{\pi}) = \frac{(\alpha_1 + m)(\alpha_2 + n - m)}{(\alpha_1 + \alpha_2 + n)^2} \tag{4.20}$$

[8] A closed-form expression is one that can be written exactly as a combination of a finite number of "well-known" functions (such as polynomials, logarithms, exponentials, and so forth).

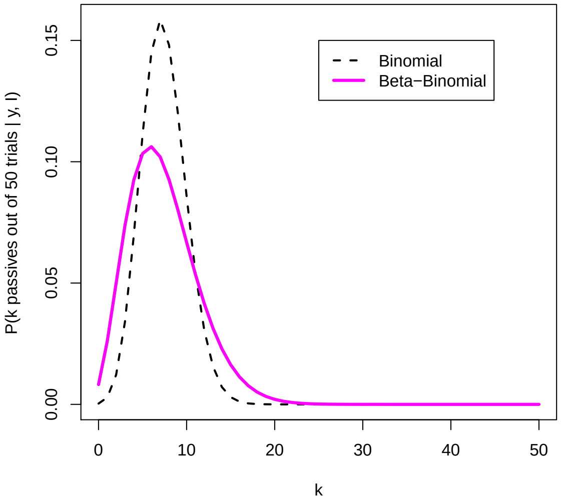


Figure 4.10: The beta-binomial model has greater dispersion than the binomial model. Results shown for $\alpha_1 + m = 5$, $\alpha_2 + n - m = 29$.

and (4.20) respectively. We were able to do this due to the CONJUGACY of the beta distribution to the binomial distribution. However, it will sometimes be the case that we want to perform Bayesian inferences but don't have conjugate distributions to work with. As a simple example, let us turn back to a case of inferring the ordering preference of an English binomial, such as {*radio, television*}. The words in this particular binomial differ in length (quantified as, for example, number of syllables), and numerous authors have suggested that a short-before-long metrical constraint is one determinant of ordering preferences for English binomials (Cooper and Ross, 1975; Pinker and Birdsong, 1979, inter alia). Our prior knowledge therefore inclines us to expect a preference for the ordering *radio and television* (abbreviated as `r`) more strongly than a preference for the ordering *television and radio* (`t`), but we may be relatively agnostic as to the particular strength of the ordering preference. A natural probabilistic model here would be the binomial distribution with success parameter π, and a natural prior might be one which is uniform within each of the ranges $0 \leq \pi \leq 0.5$ and $0.5 < \pi < 1$, but twice as large in the latter range as in the former range. This would be the following prior:

$$p(\pi = x) = \begin{cases} \frac{2}{3} & 0 \leq x \leq 0.5 \\ \frac{4}{3} & 0.5 < x \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.21}$$

which is a step function, illustrated in Figure 4.11a.

In such cases, there are typically no closed-form expressions for the posterior or predictive distributions given arbitrary observed data $\boldsymbol{y}$. However, these distributions can very often be approximated using general-purpose sampling-based approaches. Under these approaches, samples (in principle independent of one another) can be drawn over quantities that are unknown in the model. These samples can then be used in combination with density estimation techniques such as those from Chapter **??** to approximate any probability density of interest. Chapter **??** provides a brief theoretical and practical introduction to sampling techniques; here, we introduce the steps involved in sampling-based approaches as needed.

For example, suppose we obtain data $\boldsymbol{y}$ consisting of ten binomial tokens—five of `r` and five of `t`—and are interested in approximating the following distributions:

1. The posterior distribution over the success parameter π;

2. The posterior predictive distribution over the observed ordering of an eleventh token;

3. The posterior predictive distribution over the number of `r` orderings seen in ten more tokens.

We can use `BUGS`, a highly flexible language for describing and sampling from structured probabilistic models, to sample from these distributions. `BUGS` uses Gibbs sampling, a Markov-chain Monte Carlo technique (Chapter **??**), to produce samples from the posterior distributions of interest to us (such as $P(\pi|\boldsymbol{y}, I)$ or $P(\boldsymbol{y}_{new}|\boldsymbol{y}, I)$). Here is one way to describe our model in `BUGS`:

```
model {
      /* the model */
      for(i in 1:length(response)) { response[i] ~ dbern(p) }
      /* the prior */
      pA ~ dunif(0,0.5)
      pB ~ dunif(0.5,1)
      i ~ dbern(2/3)
      p <- (1 - i) * pA + i * pB
      /* predictions */
      prediction1 ~ dbern(p)
      prediction2 ~ dbin(p, 10) /* dbin() is for binomial distribution */
}
```

The first line,

```
      for(i in 1:length(response)) { response[i] ~ dbern(p) }
```

says that each observation is the outcome of a Bernoulli random variable with success parameter `p`.

The next part,

```
    pA ~ dunif(0,0.5)
    pB ~ dunif(0.5,1)
    i ~ dbern(2/3)
    p <- (1 - i) * pA + i * pB
```

is a way of encoding the step-function prior of Equation (4.21). The first two lines say that there are two random variables, pA and pB, drawn from uniform distributions on $[0, 0.5]$ and $[0.5, 1]$ respectively. The next two lines say that the success parameter p is equal to pA $\frac{2}{3}$ of the time, and is equal to pB otherwise. These four lines together encode the prior of Equation (4.21).

Finally, the last two lines say that there are two more random variables parameterized by p: a single token (prediction1) and the number of r outcomes in ten more tokens (prediction2).

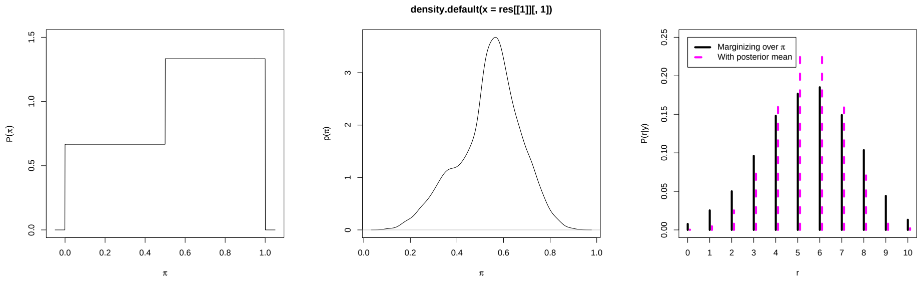
There are several incarnations of BUGS, but here we focus on a newer incarnation, JAGS, that is open-source and cross-platform. JAGS can interface with R through the R library rjags.[9] Below is a demonstration of how we can use BUGS through R to estimate the posteriors above with samples.

```
> ls()
> rm(i,p)
> set.seed(45)
> # first, set up observed data
> response <- c(rep(1,5),rep(0,5))
> # now compile the BUGS model
> m <- jags.model("../jags_examples/asymm_binomial_prior/asymm_binomial_prior.bug",dat
> # initial period of running the model to get it converged
> update(m,1000)
> # Now get samples
> res <- coda.samples(m, c("p","prediction1","prediction2"), thin = 20, n.iter=5000)
> # posterior predictions not completely consistent due to sampling noise
> print(apply(res[[1]],2,mean))
> posterior.mean <- apply(res[[1]],2,mean)

> plot(density(res[[1]][,1]),xlab=expression(pi),ylab=expression(paste("p(",pi,")")))

> # plot posterior predictive distribution 2
> preds2 <- table(res[[1]][,3])
> plot(preds2/sum(preds2),type='h',xlab="r",ylab="P(r|y)",lwd=4,ylim=c(0,0.25))
> posterior.mean.predicted.freqs <- dbinom(0:10,10,posterior.mean[1])
> x <- 0:10 + 0.1
> arrows(x, 0, x, posterior.mean.predicted.freqs,length=0,lty=2,lwd=4,col="magenta")
> legend(0,0.25,c(expression(paste("Marginizing over ",pi)),"With posterior mean"),lty
```

[9] JAGS can be obtained freely at http://calvin.iarc.fr/~martyn/software/jags/, and rjags at http://cran.r-project.org/web/packages/rjags/index.html.



(a) Prior over π

(b) Posterior over π

(c) Posterior predictive distribution for $N = 10$, marginalizing over π versus using posterior mean

Figure 4.11: A non-conjugate prior for the binomial distribution: prior distribution, posterior over π, and predictive distribution for next 10 outcomes

Two important notes on the use of sampling: first, immediately after compiling we specify a "burn-in" period of 1000 iterations to bring the Markov chain to a "steady state" with:[10]

```
update(m,1000)
```

Second, there can be AUTOCORRELATION in the Markov chain: samples near to one another in time are non-independent of one another.[11] In order to minimize the bias in the estimated probability density, we'd like to minimize this autocorrelation. We can do this by sub-sampling or "thinning" the Markov chain, in this case taking only one out of every 20 samples from the chain as specified by the argument `thin = 20` to `coda.samples()`. This reduces the autocorrelation to a minimal level. We can get a sense of how bad the autocorrelation is by taking an unthinned sample and computing the autocorrelation at a number of time lags:

```
> m <- jags.model("../jags_examples/asymm_binomial_prior/asymm_binomial_prior.bug",dat
> # initial period of running the model to get it converged
> update(m,1000)
> res <- coda.samples(m, c("p","prediction1","prediction2"), thin = 1, n.iter=5000)
> autocorr(res,lags=c(1,5,10,20,50))
```

We see that the autocorrelation is quite problematic for an unthinned chain (lag 1), but it is much better at higher lags. Thinning the chain by taking every twentieth sample is more than sufficient to bring the autocorrelation down

[10] For any given model there is no guarantee how many iterations are needed, but most of the models covered in this book are simple enough that on the order of thousands of iterations is enough.

[11] The autocorrelation of a sequence $\vec{x}$ for a time lag τ is simply the covariance between elements in the sequence that are τ steps apart, or $\mathrm{Cov}(x_i, x_{i+\tau})$.

Notably, the posterior distribution shown in Figure 4.11a looks quite different from a beta distribution. Once again the greater dispersion of Bayesian prediction marginalizing over π, as compared with the predictions derived from the posterior mean, is evident in Figure 4.11c.

Finally, we'll illustrate one more example of simple Bayesian estimation, this time of a normal distribution for the F3 formant of the vowel [æ], based on speaker means of 15 child native speakers of English from Peterson and Barney (1952). Since the normal distribution has two parameters—the mean μ and variance σ^2—we must use a slightly more complex prior of the form $P(\mu, \sigma^2)$. We will assume that these parameters are independent of one another in the prior—that is, $P(\mu, \sigma^2) = P(\mu)P(\sigma^2)$. For our prior, we choose NON-INFORMATIVE distributions (ones that give similar probability to broad ranges of the model parameters). In particular, we choose uniform distributions over μ and $\log \sigma$ over the ranges $[0, 10^5]$ and $[-100, 100]$ respectively.:[12] This gives us the model:

$$\begin{aligned} \boldsymbol{y} &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{U}(0, 10^5) \\ \log \sigma &\sim \mathcal{U}(-100, 100) \end{aligned}$$

where $\sim$ means "is distributed as".

Here is the model in BUGS:

```
var predictions[M]
model {
      /* the model */
      for(i in 1:length(response)) { response[i] ~ dnorm(mu,tau) }
      /* the prior */
      mu ~ dunif(0,100000) # based on F3 means for other vowels
      log.sigma ~ dunif(-100,100)
      sigma <- exp(log.sigma)
      tau <- 1/(sigma^2)
      /* predictions */
      for(i in 1:M) { predictions[i] ~ dnorm(mu,tau) }
}
```

The first line,

```
var predictions[M]
```

states that the `predictions` variable will be a numeric array of length `M` (with M to be specified from `R`). `BUGS` parameterizes the normal distribution differently than we have, using a precision parameter $\tau \overset{\text{def}}{=} \frac{1}{\sigma^2}$. The next line,

[12] See Gelman et al. (2004, Appendix C) for the relative merits of different choices of how to place a prior on σ^2.

```
for(i in 1:length(response)) { response[i] ~ dnorm(mu,tau) }
```

simply expresses that observations $\boldsymbol{y}$ are drawn from a normal distribution parameterized by μ and τ. The mean μ is straightforwardly parameterized with a uniform distribution over a wide range. When we set the prior over τ we do so in three stages, first saying that $\log\sigma$ is uniformly distributed:

```
log.sigma ~ dunif(-100,100)
```

and transforming from $\log\sigma$ to σ and then to τ:

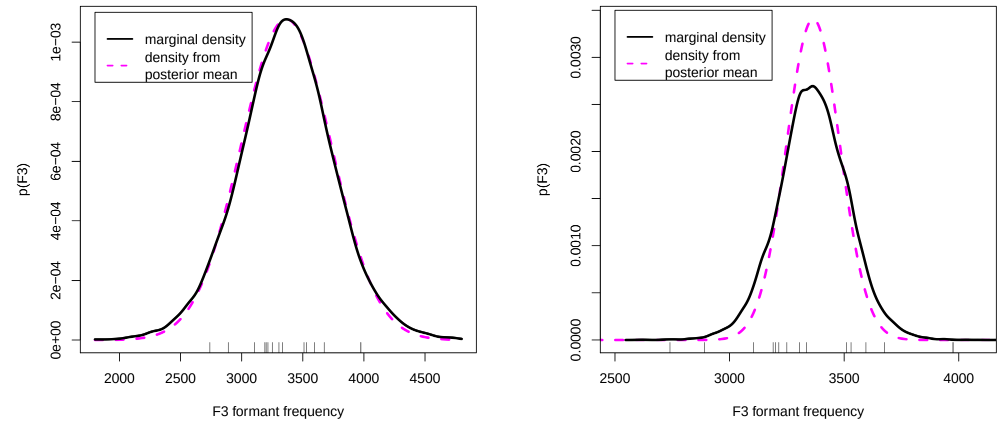
```
sigma <- exp(log.sigma)
tau <- 1/(sigma^2)
```

From R, we can compile the model and draw samples as before:

```
> pb <- read.table("../data/peterson_barney_data/pb.txt",header=T)
> pb.means <- with(pb,aggregate(data.frame(F0,F1,F2,F3), by=list(Type,Sex,Speaker,Vowe
> names(pb.means) <- c("Type","Sex","Speaker","Vowel","IPA",names(pb.means)[6:9])
> set.seed(18)
> response <- subset(pb.means,Vowel=="ae" & Type=="c")[["F3"]]
> M <- 10 # number of predictions to make
> m <- jags.model("../jags_examples/child_f3_formant/child_f3_formant.bug",data=list("
> update(m,1000)
> res <- coda.samples(m, c("mu","sigma","predictions"),n.iter=20000,thin=1)
```

and extract the relevant statistics and plot the outcome as follows:

```
> # compute posterior mean and standard deviation
> mu.mean <- mean(res[[1]][,1])
> sigma.mean <- mean(res[[1]][,12])
> # plot Bayesian density estimate
> from <- 1800
> to <- 4800
> x <- seq(from,to,by=1)
> plot(x,dnorm(x,mu.mean,sigma.mean),col="magenta",lwd=3,lty=2,type="l",xlim=c(from,to
> lines(density(res[[1]][,2],from=from,to=to),lwd=3)
> rug(response)
> legend(from,0.0011,c("marginal density","density from\nposterior mean"),lty=c(1,2),l
> # plot density estimate over mean observed in 10 more observations
> from <- 2500
> to <- 4100
> plot(x,dnorm(x,mu.mean,sigma.mean/sqrt(M)),type="l",lty=2,col="magenta",lwd=3,xlim=c
> lines(density(apply(res[[1]][,2:11],1,mean,from=from,to=to)),lwd=3) # using samples
> rug(response)
> legend(from,0.0035,c("marginal density","density from\nposterior mean"),lty=c(1,2),l
```



(a) Estimated density

(b) Density estimate over mean of ten new observations

Figure 4.12: Bayesian inference for normal distribution

The resulting density estimate for a single future observation is shown in Figure 4.12a. This is almost the same as the result obtained from using the posterior mean. However, the density estimate for the mean obtained in ten future observations, shown in Figure 4.12b, is rather different: once again it has greater dispersion than the estimate obtained using the posterior mean.[13]

The ability to specify model structures like this, drawing from a variety of distributions, and to compute approximate posterior densities with general-purpose tools, gives tremendous modeling flexibility. The only real limits are conceptual—coming up with probabilistic models that are appropriate for a given type of data—and computational—time and memory.

4.6 Further reading

Gelman et al. (2004) is probably the best reference for practical details and advice in Bayesian parameter estimation and prediction.

4.7 Exercises

Exercise 4.1

[13] The density on the mean of ten future observations under the posterior mean μ and σ^2 is given by expressing the mean as a linear combination of ten independent identically distributed normal random variables (Section 3.3).

Confirm using simulations that the variance of relative-frequency estimation of π for binomially distributed data really is $\frac{\pi(1-\pi)}{n}$: for all possible combinations of $\pi \in \{0.1, 0.2, 0.5\}, n \in \{10, 100, 1000\}$, randomly generate 1000 datasets and estimate $\hat{\pi}$ using relative frequency estimation. Plot the observed variance against the variance predicted in Equation 4.1.

Exercise 4.2: Maximum-likelihood estimation for the geometric distribution

You encountered the geometric distribution in Chapter 3, which models the generation of sequence lengths as the repeated flipping of a weighted coin until a single success is achieved. Its lone parameter is the success parameter π. Suppose that you have a set of observed sequence lengths $\boldsymbol{y} = y_1, \dots, y_n$. Since a sequence of length k corresponds to $k-1$ "failures" and one "success", the total number of "failures" in $\boldsymbol{y}$ is $\sum_i (y_i - 1)$ and the total number of "successes" is n.

1. From analogy to the binomial distribution, guess the maximum-likelihood estimate of π.

2. Is your guess of the maximum-likelihood estimate biased? You're welcome to answer this question either through mathematical analysis or through computational simulation (i.e. choose a value of π, repeatedly generate sets of geometrically-distributed sequences using your choice of π, and quantify the discrepancy between the average estimate $\hat{\pi}$ and the true value).

3. Use your estimator to find best-fit distributions for token-frequency and type-frequency distributions of word length in syllables as found in the file `brown-counts-lengths-nsyll` (parsed Brown corpus; see Exercise 3.7).

Exercise 4.3

We covered Bayesian parameter estimation for the binomial distribution where the prior distribution on the binomial success parameter π was of the form

$$P(\pi) \propto \pi^a (1-\pi)^b$$

Plot the shape of this prior for a variety of choices of a and b. What determines the mode of the distribution (i.e., the value of π where the curve's maximum lies) and its degree of peakedness? What do a and b together represent?

Exercise 4.4: "Ignorance" priors

A uniform prior distribution on the binomial parameter, $P(\pi) = 1$, is often called the "ignorance" distribution. But what is the ignorance of? Suppose we have

$$X \sim Binom(n, \pi).$$

The beta-binomial distribution over X (i.e., marginalizing over π) is $P(X = k) = \int_0^1 \binom{n}{k} \pi^n (1-\pi)^{n-k}\, d\pi$. What does this integral evaluate to (as a function of n and k) when the prior distribution on π is uniform? (Bayes, 1763; Stigler, 1986)

Exercise 4.5: Binomial and beta-binomial predictive distributions

Three native English speakers start studying a new language together. This language has flexible word order, so that sometimes the subject of the sentence can precede the verb (SV), and sometimes it can follow the verb (VS). Of the first three utterances of the new language they are taught, one is VS and the other two are SV.

Speaker A abandons her English-language preconceptions and uses the method of maximum likelihood to estimate the probability that an utterance will be SV. Speakers B and C carry over some preconceptions from English; they draw inferences regarding the SV/VS word order frequency in the language according to a beta-distributed prior, with $\alpha_1 = 8$ and $\alpha_2 = 1$ (here, SV word order counts as a “success”), which is then combined with the three utterances they’ve been exposed to thus far. Speaker B uses maximum a-posterior (MAP) probability to estimate the probability that an utterance will be SV. Speaker C is fully Bayesian and retains a full posterior distribution on the probability that an utterance will be SV.

It turns out that the first three utterances of the new language were uncharacteristic; of the next twenty-four utterances our speakers hear, sixteen of them are VS. Which of our three speakers was best prepared for this eventuality, as judged by the predictive distribution placed by the speaker on the word order outcomes of these twenty-four utterances? Which of our speakers was worst prepared? Why?

Exercise 4.6: Fitting the constituent-order model.⌨

Review the constituent-order model of Section 2.8 and the word-order-frequency data of Table 2.2.

- Consider a heuristic method for choosing the model’s parameters: set γ_1 to the relative frequency with which S precedes O, γ_2 to the relative frequency with which S precedes V, and γ_3 to the relative frequency with which V precedes O. Compute the probability distribution it places over word orders.

- Implement the likelihood function for the constituent-order model and use convex optimization software of your choice to find the maximum-likelihood estimate of $\gamma_1, \gamma_2, \gamma_3$ for Table 2.2. (In `R`, for example, the `optim()` function, using the default Nelder-Mead algorithm, will do fine.) What category probabilities does the ML-estimated model predict? How does the heuristic-method fit compare? Explain what you see.

Exercise 4.7: What level of autocorrelation is acceptable in a Markov chain?

How do you know when a given level of autocorrelation in a thinned Markov chain is acceptably low? One way of thinking about this problem is to realize that a sequence of independent samples is generally going to have *some* non-zero autocorrelation, by pure chance. The longer such a sequence, however, the lower the autocorrelation is likely to be. (Why?) Simulate a number of such sequences of length $N = 100$, drawn from a uniform distribution, and compute the 97.5% quantile autocorrelation coefficient—that is, the value r such that 97.5% of the generated sequences have correlation coefficient smaller than this value. Now repeat this process for a number of different lengths N, and plot this threshold r as a function of N.

Exercise 4.8: Autocorrelation of Markov-chain samples from BUGS.

Explore the autocorrelation of the samples obtained in the two models of Section 4.5, varying how densely you subsample the Markov chain by varying the thinning interval (specified by the `thin` argument of `coda.samples()`). Plot the average (over 20 runs) autocorrelation on each model parameter as a function of the thinning interval. For each model, how sparsely do you need to subsample the chain in order to effectively eliminate the autocorrelation? **Hint:** in R, you can compute the autocorrelation of a vector x with:

```
> cor(x[-1],x[-length(x)])
```